Attorney Docket No.: 021262-000110US

PATENT APPLICATION
METHOD AND SYSTEM FOR IDENTIFYING EXPERTISE

Inventor:

John G. Sotos, a citizen of the United States, residing at
1788 Oak Creek Drive, Apt. #415
Palo Alto, CA 94304

Entity: Small

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400

METHOD AND SYSTEM FOR IDENTIFYING EXPERTISE

COPYRIGHT NOTICE

CROSS REFERENCES TO RELATED APPLICATIONS

This application claims priority to U.S. Provisional Application No. 60/230,689, filed on September 7, 2000, entitled "METHOD AND SYSTEM FOR IDENTIFYING EXPERTISE", which is herein incorporated by reference in its entirety for all purposes.

BACKGROUND OF THE INVENTION

The present invention generally relates to knowledge acquisition. More particularly, the invention provides a method for web based searching of a database or databases in the health care field. The database can be selected from custom or commercial databases, e.g., the U.S. Government's Medline database of scientific articles (hereinafter "Medline").

Conventional ways of finding desirable hospitals and doctors are very limiting. For example, most doctors, who specialize in a selected area of practice, are often uncovered by a referral, which is generally word of mouth from one person to another person. Most often, such word of mouth is from a family member, co-worker or the like. Similarly, a desirable hospital is often found through a referral mechanism, as well. Depending upon such referral, doctor and/or hospital quality has been difficult to control. So, consumers do not presently have a reliable, objective method to find a desirable doctor and/or health care providers.

There have been some attempts to uncover experts in the healthcare field through printed documents. Such printed documents include, among others, periodicals, etc. To identify such experts in a field who may be able to comment on breaking health news

stories, consumers must often conduct extensive research or use limited, non-objective printed guides. Similarly, other professionals in the field such as the media also need to perform extensive research and the like to find such breaking heath news stories. Still further, makers of healthcare policy and reimbursement programs desire methods to identify centers of excellence around which specialized health care services could be consolidated. Accordingly, it is often difficult to uncover such experts in the healthcare field using conventional techniques.

From the above, it is seen that a technique for improving access to information is highly desirable.

SUMMARY OF THE INVENTION

According to the present invention, a technique including a method and system for identifying an entity with an expertise is provided. In an exemplary example, the invention provides a method for web based searching of a database or databases in the health care field. The database can be selected from custom or commercial databases, e.g., Medline, etc.

In a specific embodiment, the invention provides a method of identifying entities having expertise in one or more subjects. Among other features, the method includes querying a database for documents (e.g., articles, papers, periodicals) relevant to a subject; and calculating a first score for each relevant document. The method also identifies entities affiliated with one or more relevant documents; and calculates a second score for each entity based on the one or more first scores of the one or more relevant documents affiliated with the entity. A step of ranking expertise of the entities based on the respective second scores of the entities is included.

In an alternative embodiment, the invention provides a system for identifying expertise in a subject. The system has a server system coupled with a database of documents. The server system has a memory with a variety of computer codes. The codes include a first code configured to receive a subject from a user; and a second code configured to query the database for documents relevant to the subject. The memory also has a third code configured to calculate a score for each relevant document; and a fourth code configured to identify entities affiliated with one or more of the relevant documents. A fifth code is configured to calculate a score for each entity based on the one or more scores of the one or more relevant documents affiliated with the entity. A sixth code is configured to provide to the user a ranking of the expertise of the entities based on the respective scores of the entities.

Depending upon the embodiment, the invention can also include only some of the codes above, as well as other codes, which can implement the functionality described herein.

Numerous benefits are achieved by way of the present invention. In a specific embodiment, the invention can be implemented using conventional hardware and/or software. In other aspects, the invention can be used to find a leading expert in a particular health care field. The invention can also be used to find a leading hospital for a particular health care field. Still further, the invention can be used to uncover almost any entity, which is desirable for a particular field. Geographical areas can also be uncovered for particular expertise. In most embodiments, the invention is easy to use and can be implemented on a computer network, such as a local area network, a wide area network, a world wide area network, any combination of these, and the like. Depending upon the embodiment, one or more of these benefits may be achieved. These and other benefits are provided in more detail throughout the present specification and more particularly below.

These and other embodiments of the present invention, as well as its advantages and features are described in more detail in conjunction with the text below and attached.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a simplified block diagram of an environment in which methods according to various embodiments of the invention may be implemented;

Fig. 2 is a simplified block diagram of basic subsystems in a representative computer system in which methods according to various embodiments of the invention may be implemented;

Fig. 3 is a simplified flow diagram of a method according to one embodiment of the invention; and

Figs. 4-20 illustrate examples of web pages that may be used to implement methods according to various embodiments of the invention

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

System Overview

Fig. 1 is a simplified block diagram of an environment in which methods according to various embodiments of the invention may be implemented. This diagram is merely for illustrative purposes and is not intended to limit the scope of the claims herein. An expertise finder server system 102 is coupled with a database 104 and a network 106.

Additionally, a database server system 108 is coupled with a database 110 of documents and the network 106. Additionally, a user computer 112 is coupled with the network 106.

Network 106 allows the expertise finder server system 102, the database server system 108, and the user computer 112 to communicate with other computers and with each other. The network 106 may be a local area network, a wide area network, an internet, the Internet, an intranet, an extranet, or the like.

The expertise server system 102 may be coupled to the network 106 via a relatively high bandwidth transmission medium such as a T1 or T3 line. With respect to the electronic database 104, it generally contains WebPages, forms, etc. The database 104 can be composed of a number of different databases. These databases can be located in one central repository, or alternatively, they can be dispersed among various distinct physical locations. These databases can be categorized and structured in various ways based on the needs and criteria of the database designer. Methods used to create and organize databases are commonly known in the art, for example, relational database techniques can be used to logically connect these databases. Database 104 can be a relational database, distributed database, object-oriented database, mixed object-oriented database, or the like. Database products with which the present invention may be implemented include, but are not limited to, IBM's DB2, Microsoft's Access and FoxPro, and database products from Oracle, Sybase, and Computer Associates.

In one embodiment, the database or databases 104 can be physically located separate from the expertise finder server system 102. These databases can reside on remote, distant servers on a local area network or the Internet. Under this arrangement, whenever any data are needed, the processor needs to access the necessary database(s) via a communication channel to retrieve the requisite data for analysis. For example, the processor can access and retrieve data from a remote database via a computer network such as a LAN or the Internet. Generally, the expertise server system 102 and database 104 store information and disseminate it to individual computers *e.g.* 112 over the network 106.

The database server system 108 may also be coupled to the network 106 via a relatively high bandwidth transmission medium such as a T1 or T3 line. With respect to the electronic database 110, it generally contains electronic versions of documents such as, for example, technical articles and papers, in fields such as medicine, genetics, physics, chemistry, engineering, law, and the like. The database 110 can also include legal cases. As discussed with respect to the database 104, the database 110 can comprise a number of different databases. In one embodiment, the database 110 is the Medline database. It is to be

understood that other databases may be used instead of, or in conjunction with, Medline, such as other Medlars databases, the Science Citation Index, Medlex, Westlaw, Lexis, Dialog, and the like.

In some embodiments, the expertise server system 102 and the database server system 108 are the same server system, whereas in other embodiments, the two server systems are separate. Similarly, in some embodiments, the database 104 and the document database 110 are included in one database system, whereas in other embodiments, the two databases are separate.

The user computer 112 may be configured with many different hardware components and can be made in many dimensions, styles and locations (*e.g.*, laptop, palmtop, pentop, server, workstation and mainframe). The user computer 112 may be, as one example, a conventional desktop personal computer or workstation having the ability to connect to network 106 and being capable of running customized software supporting the service provided by the present invention. In some embodiments, user computer 112 includes web browsing software, or the like for interacting with expertise finder server system 102. In a specific embodiment, the network 106 is the Internet, and the expertise server system 102 provides Web Pages to computers, such as user computer 112, via the Internet 106.

Embodiments according to the present invention can be implemented in a single application program, or can be implemented as multiple programs in a distributed computing environment, such as a workstation, personal computer or a remote terminal in a client server relationship. Fig. 2 illustrates basic subsystems in a representative computer system in which methods according to various embodiments of the invention may be implemented. In these embodiments, each of user computers 112, expertise finder server system 102, and/or database server system 108 may comprise one or more of computer systems 200 or the like. This diagram is merely an illustration and should not limit the scope of the claims herein. One of ordinary skill in the art will recognize other variations, modifications, and alternatives

In certain embodiments, the subsystems are interconnected via a system bus 202. Additional subsystems such as a printer, keyboard, fixed disk and others are shown. Peripherals and input/output (I/O) devices can be connected to the computer system by any number of means known in the art, such as serial port 204. For example, serial port 204 can be used to connect the computer system to a modem, which in turn connects to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus 202 allows central processor 206 to communicate with each subsystem and to

control the execution of instructions from system memory 208 or the fixed disk, as well as the exchange of information between subsystems. Other arrangements of subsystems and interconnections are readily achievable by those of ordinary skill in the art. System memory 208, and the fixed disk are examples of tangible media for storage of computer programs, other types of tangible media include floppy disks, removable hard disks, optical storage media such as CD-ROMS and bar codes, and semiconductor memories such as flash memory, read-only-memories (ROM), and battery backed memory.

Finding Expertise

The invention will now be discussed in the context of medical expertise. However, it is to be understood that the present invention is not limited to identifying medical expertise. On the contrary, the present invention may be used to identify expertise in many different fields such as healthcare, genetics, physics, chemistry, engineering, law, and the like.

Fig. 3 is a simplified flow diagram of a method according to one embodiment of the invention. This diagram is merely for illustrative purposes and is not intended to limit the scope of the claims herein. The flow illustrated in Fig. 3 may be implemented, for example, on a system such as the system 100 shown in Fig. 1. In a step 252, a subject for which a user desires to identify expertise is received. In some embodiments, the subject may be received via a world wide network of computers such as the Internet, or the like. Referring to Fig. 1, in a specific embodiment, the user computer 112 transfers a subject to the expertise server system 102 via the Internet.

Then, in a step 254, a database is queried to find documents relevant to the subject. In a specific embodiment, the expertise server system 102 queries the database 110 via the Internet and via the database server system 108.

After receiving a response to the database query, a score is calculated, in a step 256, for each document identified as relevant to the subject. In a specific embodiment in which the Medline database is queried, a document that is a review article is given more points than an article that is not a review article. Additionally, different types of review articles are awarded different point amounts.

Then, in a step 258, a score is calculated for each author identified as an author of one or more of the relevant documents. In one embodiment, the score of an author is based on the scores of documents authored by the author. In another embodiment, the

author's score is additionally based on, for each document, the position of the author's name with respect to other authors.

In a step 260, a score may be calculated for institutions affiliated with the documents. In one embodiment, a score for an institution is calculated based on the scores of documents that emanate from the institution. In another embodiment, the score is based on the scores of authors affiliated with the institution. Similarly, in a step 262, a score may be calculated for geographic areas. In one embodiment, a score for a geographic area is calculated based on the scores of documents that emanate from the institutions in the geographical area. In another embodiment, the score is based on the scores of authors associated with the geographic area. Then, in a step 264, the scores of one or more of the documents, authors, institutions, and geographic areas are ranked and the one or more corresponding rankings are provided to the user.

One skilled in the art will recognize many variations, modifications, and alternatives to the flow of Fig. 3. For instance, in some embodiments only one of steps 258, 260 and 262 are performed. In other embodiments, any two of steps 258, 260 and 262 are performed.

Querying the Database

As described above, a database of documents is queried for documents relevant to a subject of interest. In some embodiments, the query is determined such that documents relevant to the subject of interest, but not necessarily relevant to determining expertise, will be excluded. For example, in the context of finding medical expertise, the query could be determined such that letters to the editor will be excluded. Also, in the context of finding clinical care expertise, the query could be determined such that documents that do not deal with humans will be excluded.

In some embodiments, the query is determined such that synonyms of the subject provided by the user are also included in the query. For example, a user may provide the subject "kidney disease," and a query could be determined that included "renal disease." In a specific embodiment, a database of synonyms is coupled to expertise finder server system 102. The database of synonyms can be included in database 104, or in a separate database. In this embodiment, upon receiving a subject from the user, the database of synonyms is queried to determine whether any synonyms should be included in the query of document database 110.

In one specific embodiment, the query of document database 110 limits documents to those from the most recent ten years. In some other embodiments, the query may limit to some other number of years (e.g., 5, 15, etc.). In yet other embodiments, the query does not include such a time limit.

One skilled in the art will recognize many other variations and alternatives. For instance, queries can be optimized based on the particular subject. For example, a query for a search for expertise related to the diagnosis of disease can be determined according to one optimization, whereas a query for a search for surgical expertise can be determined according to another optimization.

Scoring Documents

In embodiments according the invention, scores for documents are calculated based on information transmitted from the database server 108 to the expertise finder server 102 in response to a query. In some embodiments, the expertise finder server 102 receives a document record for each of one or more documents in response to the query. Each document record includes one or more data fields. For example, the Medline database includes the following fields: TITLE, AUTHOR, ADDRESS, DATE PUBLISHED, and PUBLICATION TYPE.

In these embodiments, a score for a document is calculated based on information returned in the data fields. For example, a point value can be assigned to each of a plurality of data fields for a document, and then the point values summed to calculate the score for the document. The point value assigned to a particular data field can be determined by the information returned in that data field.

For instance, in some embodiments, a data field may indicate the type of document, and a score assigned to the type of document based on the information provided in this data field. In a specific embodiment, a higher point value is assigned to review articles than non-review articles. And, in another specific embodiment, some types of review articles are assigned point values than other types of review articles.

One particular simplified example of an algorithm for scoring a document from the Medline database is provided. In this example, if the PUBLICATION TYPE data field has value REVIEW or PRACTICE GUIDELINE, the document is given 10 points. If the PUBLICATION TYPE data field has value LETTER, the document is assigned 0 points. Otherwise, the document is assigned 1 point. Additionally, the document's score can be adjusted according to when it was published. For example, if the DATE PUBLISHED field

indicates that the document was published within the last 3 years, then the document score is increased by 5 points. The above example is merely an illustration of one of the many ways in which a document may be scored. Additionally, further details are provided in the attached Appendix. One skilled in the art will recognize many other equivalents, modifications, and alternatives. For example, the same scoring algorithm need not be used for each search. Different scoring algorithm can be used, for example, according to the particular database queried, the particular subject, the number of document records returned by the query, etc.

In other embodiments, some of the optimizations related to the query described above can be implemented in the scoring algorithm rather than in determining the query. For example, a letter to the editor, a paper that does not consider humans, etc. can be assigned a score of zero.

In some embodiments, a scoring algorithm can vary according to the number of documents that were returned by the query. For example, in the context of finding medical expertise, it has been found that when only a few documents are returned, the type of document (review article or whatever) is less informative. Thus, when the number of documents returned are low, the data field related to the type of document can be assigned a same point value, regardless of the information returned in this data field. In one specific example, the point value assigned is zero.

In other embodiments, the score of a document can be adjusted based on the type of publication in which it appeared. For instance, some publications, journals, etc., may be more highly regarded than others. This has been quantified in various ways. For example, the Institute for Scientific Information assigns an "impact factor" to scientific journals. Thus, in some embodiments, the "impact factor" assigned to the journal can be used to adjust the score of the document.

Scoring Authors

In some embodiments according to the invention, scores for authors are calculated. In some of these embodiments, a score for an author is calculated based on the score of each document returned of which he/she is an author. In one specific embodiment, the score for an author is the sum of the scores of the documents of which he/she is an author. In other embodiments, the score for an author is a weighted sum of the scores of the documents of which he/she is an author. For example, a higher weight may be assigned to a document if the author is the first listed than when the author is not the first listed. Similarly,

in some embodiments, a higher weight may be assigned to a document if the author is the "corresponding author" (i.e. the author to whom correspondence is addressed). In other embodiments, an author's score may be increased by an increment for each document of which he/she is first listed or a corresponding author. In one specific embodiment in which the Medline database is queried, the corresponding author of a document is determined by examining the email address often attached to a document's ADDRESS field.

Scoring Institutions/Geographic Areas

In some embodiments according to the invention, scores for institutions are calculated. In some of these embodiments, a score for an institution is calculated based on the score of each document that emanated from that institution. In one specific embodiment, the score for an institution is the sum of the scores of the documents that emanated from that institution. In other embodiments, the score for an institution is a weighted sum of the scores of the documents that emanated from that institution.

Some databases may include a data field that indicates from which institution or institutions the document emanated. In other databases, such a field may not be included. For example, the Medline database sometimes, but not always, provides the institutional affiliation of the first author in the ADDRESS field. Thus, one of the institutions from which the document emanated can be assumed to be the institution found in the ADDRESS field.

With the Medline database, there is no standard nomenclature for institutions and there is no standard syntax for the contents of the ADDRESS field. Thus, in one particular embodiment in which the Medline database is searched, a plurality of heuristics are employed to parse the contents of the ADDRESS field and compare to a database of institutions.

The above-described embodiment provides a method for determining the institution with which the first author of a document is affiliated. In one particular embodiment, it is assumed that the other authors (if any) are affiliated with the same institution. In other embodiments, determining the one or more institutions from which a document emanated can include heuristics that examine, for example, any or all of: (a) clustering of authors on a plurality of papers, (b) institutional affiliations of the author when the author is the first author, and (c) temporal variations in imputed institutional affiliations.

Similarly, in some embodiments according to the invention, scores for geographic areas are calculated. In some of these embodiments, a score for a geographic area is calculated based on the score of each document that emanated from that geographic area.

In one specific embodiment, the score for a geographic area is the sum of the scores of the documents that emanated from that geographic area. In other embodiments, the score for a geographic area is a weighted sum of the scores of the documents that emanated from that geographic area.

In some embodiments, from which geographic region(s) a particular document emanated can be determined based on the determination of the institution(s) from which the document emanated. For example, if it is determined that a particular document emanated from Johns Hopkins, then it can be determined that document emanated from Baltimore, Maryland. In other embodiments, from which geographic region(s) a particular document emanated can be determined by examination of data fields returned by the query. For example, in embodiments that query the Medline database, the ADDRESS field can be examined to determine a geographic area.

In some embodiments, the score of an institution or geographic area is determined based on the scores of the authors affiliated with the institution or geographic area. In a specific embodiment, the determination of the score of an institution or geographic area includes employing a normalization factor that modifies the effect of having multiple authors on the same document from the same institution or geographic region.

In one specific embodiment in which the Medline database is queried, an ADDRESS field that is returned for a document is parsed into a structured geographic representation. Unlike most of the Medline data fields, the ADDRESS field is natural language text. In this specific embodiment, tokenizing rules and heuristics, as well as a database, are employed to identify institutions and geographical entities within the field. For example, an email address is examined and it is determined that the author is from Greece when the author's email address ends in ".gr". Similarly, it is determined that the author of a document is affiliated with Harvard University when "Children's Hospital" appears in a Boston address line. Also, it is determined that the author of a document is affiliated with the University of Pennsylvania when "Children's Hospital" appears in a Philadelphia address.

Displaying Rankings

In some embodiments, expertise scores of authors, institutions, geographic areas are represented with score bars. Depending upon the particular algorithm used for calculating expertise scores, the point scale may be arbitrary. Thus, the absolute numbers of the expertise scores are less useful than the relative scores. In some embodiments, expertise scores are provided to the user as bars. For example, the longer the bar, the greater the

expertise. In a specific embodiment, a normalizing bar is provided on the display of expertise scores. For example, it has been discovered that Palo Alto is the world's city with the greatest expertise in sleep apnea. So, if a user desires to view the levels of expertise for various cities in Ohio, for example, the Palo Alto scoring bar is additionally displayed, so that the user can compare how the Ohio expertise scores compare with the highest score, Palo Alto.

User Interface

Figs. 4-20 illustrate examples of web pages that can be used to implement one embodiment according to the invention. For instance, Figs. 4-6 illustrate an example of a web page form 302 for providing a subject for which a user desires to identify expertise. The form 302 includes user interface components well known to those skilled in the art, including type-in boxes 304 and 306, selectable menus 308 and 310, and submit buttons 312 and 314. Examples of drop-down menus are illustrated in Figs. 5 and 6.

Fig. 7 and Figs. 12-20 illustrate various examples of displaying expertise rankings to a user. Figs. 8-11 illustrate examples of drop-down menus for selecting display-options for expertise rankings. The examples illustrate one embodiment that provides a user interface that permits a user to choose a desired level of geographic detail.

Variations

In some embodiments, when a document cannot be classified in full geographical detail (down to the level of institution), it is assigned into an "unclassified" bucket. For example, parsing the document's ADDRESS field may show it emanated from Cleveland, but the institution is unknown. The document is classified as "unclassified Cleveland". As another example, if a document emanated from Michigan, but the city and institution are unknown, the document is classified as "unclassified Michigan".

In some embodiments, a particular person, institution, etc., is determined as affiliated with a document if there is an exact match of the name of the person, institution, etc., with the document (e.g., within a data field returned in response to a query). In other embodiments, such a determination can be made also when the name is not an exact match, but similar (e.g., J. Sotos and J.G. Sotos). In a particular embodiment, rules for determining whether similar names refer to the same author, institution, etc., takes into account that common names, such as, for example, J Smith and JG Smith, can less reliably be assumed the same.

Although the Medline ADDRESS field is supposed to contain data about the first author of a paper, sometimes another author's email address is given in the ADDRESS field. In the future, it would make sense for the system to try matching the email address in the ADDRESS field with the author to which it properly belongs.

To prove the principle and operation of the present invention, we implemented aspects of the invention using computer source code. As merely an example, the computer source code is provided in the Appendix, which is incorporated by reference. The source code was prepared in HTML/Javascript and Macintosh Common Lisp., which is not intended to be limiting in any manner. Here, one of ordinary skill in the art would recognize many other variations, alternatives and modifications. The present code can be implemented in the hardware described herein, as well as others. It could also be distributed or even implemented solely in hardware. The functionality of the source code can be further combined or even separated.

Numerous benefits are achieved by way of the present invention. Embodiments of the invention can be implemented using conventional hardware and/or software. In other aspects, embodiments of the invention can be used to find a leading expert in a particular health care field. Embodiments of the invention can also be used to find a leading hospital for a particular health care field. Still further, embodiments of the invention can be used to uncover almost any entity, which is desirable for a particular field. For example, embodiments of the present invention can be used to find testifying experts for trials. Geographical areas can also be uncovered for particular expertise. In most embodiments, the invention is easy to use and can be implemented on a computer network, such as a local area network, a wide area network, a world wide area network, any combination of these, and the like. Depending upon the embodiment, one or more of these benefits may be achieved.

Although the above has been described in terms of a healthcare database, many variations can exist. For example, the invention can be implemented using a financial database, an educational database, a legal database, a scientific database, a travel database, a entertainment database, a database of books and the like, a mating database, among others. One of ordinary skill in the art would recognize many other variations, modifications, and alternatives.

While the above is a full description of the specific embodiments, various modifications, alternative constructions and equivalents may be used. Therefore, the above

description and illustrations should not be taken as limiting the scope of the present invention which is defined by the appended claims.